On p-Values and Bayes Factors

Held, Leonhard ; Ott, Manuela

Abstract: The p-value quantifies the discrepancy between the data and a null hypothesis of interest, usually the assumption of no difference or no effect. A Bayesian approach allows the calibration of p-values by transforming them to direct measures of the evidence against the null hypothesis, so-called Bayes factors. We review the available literature in this area and consider two-sided significance tests for a point null hypothesis in more detail. We distinguish simple from local alternative hypotheses and contrast traditional Bayes factors based on the data with Bayes factors based on p-values or test statistics. A well-known finding is that the minimum Bayes factor, the smallest possible Bayes factor within a certain class of alternative hypotheses, provides less evidence against the null hypothesis than the corresponding p-value might suggest. It is less known that the relationship between p-values and minimum Bayes factors also depends on the sample size and on the dimension of the parameter of interest. We illustrate the transformation of p-values to minimum Bayes factors with two examples from clinical research.

DOI: https://doi.org/10.1146/annurev-statistics-031017-100307

On P-values and Bayes factors

Leonhard Held and Manuela Ott

Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland, CH-8001; email: leonhard.held@uzh.ch, manuela.ott@uzh.ch

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–28

https://doi.org/10.1146/((please add article doi))

Keywords

Bayes factor, evidence, minimum Bayes factor, objective Bayes, P-value, sample size

Abstract

The P-value quantifies the discrepancy between the data and a null hypothesis of interest, usually the assumption of no difference or no effect. A Bayesian approach allows to calibrate P-values by transforming them to direct measures of the evidence against the null hypothesis, so-called Bayes factors. We review the available literature in this area and consider two-sided significance tests for a point null hypothesis in more detail. We distinguish simple from local alternative hypotheses and contrast traditional Bayes factors based on the data with Bayes factors based on P-values or test statistics. A well-known finding is that the minimum Bayes factor, the smallest possible Bayes factor within a certain class of alternative hypotheses, provides less evidence against the null hypothesis than the corresponding P-value might suggest. It is less known that the relationship between P-values and minimum Bayes factors also depends on the sample size and on the dimension of the parameter of interest. We illustrate the transformation of P-values to minimum Bayes factors with two examples from clinical research.

1. INTRODUCTION

P-value: the probability, under the assumption of no effect (the null hypothesis H_0), of obtaining a result equal to or more extreme than what was actually observed.

The P-value is the probability, under the assumption of no association or no effect (the null hypothesis H_0), of obtaining a result equal to or more extreme than what was actually observed (Goodman 2005). P-values for point null hypotheses still dominate most of the applied literature (Greenland & Poole 2013), despite the fact that P-values are commonly misused (Wasserstein & Lazar 2016; Matthews et al. 2017). Specifically, a quantitative interpretation of P-values beyond the notorious dichotomization into "significant" and "non-significant" has caused a lot of confusion and misinterpretations are commonplace. Most prominent is the widespread belief that the P-value is the probability of a "chance finding", *i. e.* the probability of the null hypothesis, but many other misinterpretations can also be found (Goodman 2008; Greenland et al. 2016).

A first step towards a quantitative interpretation of P-values is a categorization into more than two levels, making a step away from the Neyman-Pearson hypothesis test paradigm to Fisher's significance test. Cox & Donnelly (2011, Section 8.4) give the following guidelines to interpret P-values as measures of evidence against a null hypothesis H_0: if $p \simeq 0.1$ there is "a suggestion of evidence" against H_0; if $p \simeq 0.05$ there is "modest evidence" against H_0; if $p \simeq 0.01$ there is "strong evidence" against H_0. Bland (2015, Section 9.4) suggests a similar "rough and ready guide" with five levels, reproduced in Table 1.[1] Similar categories have been proposed in many other applied statistics textbooks, for example Ramsey & Schafer (2002).

However, such categorizations always carry a level of arbitrariness. In addition, P-values are only indirect measures of evidence: A P-value is computed under the assumption that the null hypothesis H_0 is true, so it is conditional on H_0. It does not allow for conclusions about the probability of H_0 given the data, which is usually of primary interest. More precisely, a P-value is a quantitative measure of discrepancy between the data and the point null hypothesis H_0 (Goodman 1999a). But, as Cox (2006, page 83) puts it, "conclusions expressed in terms of probability are on the face of it more powerful than those expressed indirectly via confidence intervals and p-values". Such direct conclusions can be obtained by using Bayes factors. Assuming an alternative hypothesis H_1 has also been specified, the Bayes factor directly quantifies whether the data have increased or decreased the odds of H_0. A better approach than categorizing a P-value is thus to transform a P-value to a Bayes factor or a lower bound on a Bayes factor, a so-called minimum Bayes factor (Goodman 1999b). But many such ways have been proposed to calibrate P-values, and there is currently no consensus how P-values should be transformed to Bayes factors.

One-sided *P*-value: based on the probabilities of extreme values in one pre-specified direction of a point null hypothesis.

First, there is an important distinction between tests for direction and tests for existence (Marsman & Wagenmakers 2017). Tests for direction investigate whether the parameter of interest is above or below a specific value, assuming that there is an effect. For example, a test for direction can be used to assess whether a treatment effect is positive or negative. Tests for direction are usually conducted with one-sided P-values and there is a close correspondence to the Bayesian approach based on the posterior probability that the effect is positive or negative. In fact, this posterior probability is often equal or approximately equal to the one-sided P-value, if a non-informative prior is used (Casella & Berger 1987). A simple example is given in Lee (2004, Section 4.2).

[1] Note that the categories in the right column are shifted since Cox & Donnelly (2011) specify the amount of evidence of specific P-values ($p \simeq 0.1$, 0.05 and 0.01), which correspond to certain cutpoints in the categorization by Bland (2015).

	Strength of evidence against H_0	
P-value	Bland (2015)	Cox & Donnelly (2011)
> 0.1	Little or no evidence	A suggestion of evidence
0.1 to 0.05	Weak evidence	Modest evidence
0.05 to 0.01	Evidence	Strong evidence
0.01 to 0.001	Strong evidence	(not available)
< 0.001	Very strong evidence	

Table 1 Categorization of P-values into levels of evidence against H_0

In contrast, tests for existence want to summarize the evidence against the point null hypothesis of no effect. Tests for existence can be conducted with one-sided or two-sided P-values, but the correspondence of the P-value to the Bayesian posterior probability of the null is now lost and care has to be taken to transform P-values to Bayes factors.

Two-sided P-value: based on the probabilities of extreme values in both directions of a point null hypothesis.

In this paper we consider tests for existence. We will review different methods being proposed to calibrate P-values, identify problems with some of the proposed methods and give general recommendations how to transform P-values to (minimum) Bayes factors. We will emphasize that this transformation depends on how the P-value has been calculated. Specifically, the sample size as well as the dimension of the parameter of interest matters. It also matters whether the P-value came from a study with a well-defined alternative hypothesis, or from a study used to generate possible hypotheses.

1.1. Bayes Factors

Consider a significance test for existence with a point null hypothesis H_0: $\theta = \theta_0$ where the parameter of interest θ may be a scalar or a vector. In many problems $\theta_0 = 0$, for example when testing if there is evidence for a difference θ between two treatment groups. The alternative hypothesis may be simple, *i. e.* H_1: $\theta = \theta_1 \neq \theta_0$ or composite, usually H_1: $\theta \neq \theta_0$. In the latter case, a Bayesian approach now requires a prior distribution $f(\theta \,|\, H_1)$ to be specified. Local alternatives, represented by a unimodal symmetric prior distribution centered around the null value θ_0, are the common choice. In contrast, non-local alternatives (Johnson & Rossell 2010) have zero probability mass in a neighborhood of θ_0, with the simple alternative H_1: $\theta = \theta_1 \neq \theta_0$ being a special case.

Local alternatives: a unimodal symmetric prior distribution of alternatives centered around the null value.

The Bayes factor (BF) transforms the prior odds $\mathsf{Pr}(H_0)/\mathsf{Pr}(H_1)$ (where $\mathsf{Pr}(H_1) = 1 - \mathsf{Pr}(H_0)$) to the posterior odds $\mathsf{Pr}(H_0 \,|\, y)/\mathsf{Pr}(H_1 \,|\, y)$ in the light of the data y:

$$\frac{\mathsf{Pr}(H_0 \,|\, y)}{\mathsf{Pr}(H_1 \,|\, y)} = \mathrm{BF}(y) \cdot \frac{\mathsf{Pr}(H_0)}{\mathsf{Pr}(H_1)}. \tag{1}$$

Bayes factor: compares the likelihood of the data y under the null hypothesis H_0 to the likelihood under the alternative hypothesis H_1.

The Bayes factor $\mathrm{BF}(y)$ thus is a direct quantitative measure how the data y have increased or decreased the odds of H_0, regardless of the actual value of the prior probability $\mathsf{Pr}(H_0)$. The Bayes factor (or its logarithm) is therefore often referred to as the "strength of evidence" or "weight of evidence" (Good 1950; Bernardo & Smith 2000). If necessary, we may add an index to $\mathrm{BF}(y)$, where $\mathrm{BF}_{01}(y)$ stands for "H_0 versus H_1", so $\mathrm{BF}_{10}(y) = 1/\mathrm{BF}_{01}(y)$.

In (1), the Bayes factor

$$\mathrm{BF}(y) = \frac{f(y \,|\, H_0)}{f(y \,|\, H_1)} \tag{2}$$

	Strength of evidence against H_0		
Bayes factor	Jeffreys (1961)	Goodman (1999b)	Held & Ott (2016)
1 to 1/3	Bare mention	Weak	Weak
1/3 to 1/10	Substantial	Moderate	Moderate
1/10 to 1/30	Strong	Moderate to strong	Substantial
1/30 to 1/100	Very strong	Strong to very strong	Strong
1/100 to 1/300	Decisive	(not available)	Very strong
$< 1/300$			Decisive

Table 2 Categorization of Bayes factors BF ≤ 1 into levels of evidence against H_0

is the ratio of the likelihood $f(y \mid H_0) = f(y \mid \theta = \theta_0)$ of the observed data y under the null hypothesis H_0 and the likelihood

$$f(y \mid H_1) = \int f(y \mid \theta) f(\theta \mid H_1) d\theta \tag{3}$$

Marginal likelihood: the average likelihood with respect to a prior distribution for alternative hypotheses.

under the alternative hypothesis H_1. For a simple alternative, (3) reduces to the ordinary likelihood $f(y \mid H_1) = f(y \mid \theta = \theta_1)$ and the Bayes factor (2) reduces to a likelihood ratio. In general (3) represents a marginal likelihood, *i. e.* the average likelihood $f(y \mid \theta)$ with respect to the prior distribution $f(\theta \mid H_1)$ for θ under the alternative H_1 (Kass & Raftery 1995). Note that the computation of the Bayes factor via (2) does not require the specification of the prior probability $\Pr(H_0)$.

In this paper we focus on the evidence against a point null hypothesis provided by small Bayes factors $\text{BF}_{01} \leq 1$, such that Bayes factors lie in the same range as P-values, which facilitates comparisons. To categorize such Bayes factors, Held & Ott (2016) provided a six-grade scale reproduced in Table 2, which was proposed as a compromise of the grades proposed in Jeffreys (1961, Appendix B) and Goodman (1999b, Table 1 and 2) (also shown in Table 2). [2]

Communication of Bayes factors is of central importance. The categories shown in Table 2 are helpful in this respect, but there remains a level of arbitrariness in the definition of the category levels. Ideally, the Bayes factor itself should be reported and comprehensive formatting of Bayes factors is now crucial. We recommend to present Bayes factors as ratios, for example $\text{BF}_{01} = 1/7$, since this underlines the symmetry of Bayes factors if numerator and denominator are exchanged, here $\text{BF}_{10} = 7/1$. For Bayes factors smaller than 1/10, say, it is usually sufficient to report Bayes factors in the $1/x$ format, where x is an integer. If the Bayes factor is larger, then we recommend to use an additional decimal place for x, e.g. BF= 1/2.5 or BF= 1/1.3, to achieve better accuracy.

The Bayes factor (2) is based on the data y, sometimes called a data-based Bayes factor (Held et al. 2015) to distinguish it from Bayes factors based on test statistics or P-values. Indeed, the step from a P-value p to a Bayes factor is most easily accomplished by treating p as the data y in (2) to obtain a P-based Bayes factor based on the sampling distribution

[2] Jeffreys has actually used the slightly different cutpoints $(1/\sqrt{10})^a$, $a = 1, 2, 3, 4$, whereas Goodman has specified his evidence categories for Bayes factors of 1/5, 1/10, 1/20 and 1/100, which we have somewhat shifted to our cutpoints 1/3, 1/10, 1/30 and 1/100.

of p under H_0 and H_1:

$$\mathrm{BF}(p) = \frac{f(p \,|\, H_0)}{f(p \,|\, H_1)}. \tag{4}$$

P-based Bayes factor: a Bayes factor that is based on the sampling distributions of the P-value.

The distribution $f(p \,|\, H_0)$ of a two-sided P-value p under H_0 can usually be assumed to be uniform, since the corresponding Neyman-Pearson hypothesis test is constructed to maintain any Type-I error rate α, *i. e.* $\mathsf{Pr}(p \leq \alpha \,|\, H_0) = \alpha$ for all $\alpha \in (0, 1)$ and so $f(p \,|\, H_0) = 1$ for all p and therefore $\mathrm{BF}(p) = 1/f(p \,|\, H_1)$. The distribution $f(p \,|\, H_1)$ will depend on the specific problem considered, see Hung et al. (1997) for a comprehensive study and Donahue (1999) for a specific example. A simple option is to directly specify a distribution for $p \,|\, H_1$, for example a beta distribution. P-based Bayes factors are particularly useful if only the P-value, but not the underlying data is available.

The other option is to back-transform p to the underlying test statistic t, which was used to calculate p. If this transformation is one-to-one, then $t = t(p)$ is well defined and it is easy to see that the Bayes factor does not change, if we use t rather than p:

$$\mathrm{BF}(t) = \frac{f_t(t(p) \,|\, H_0)}{f_t(t(p) \,|\, H_1)} = \frac{f_p(p \,|\, H_0)}{f_p(p \,|\, H_1)} = \mathrm{BF}(p), \tag{5}$$

Test-based Bayes factor: a Bayes factor that is based on the sampling distributions of a test statistic.

since $f_t(t(p) \,|\, H_i) = f_p(p \,|\, H_i)\,|dt(p)/dp|^{-1}$ for $i = 0, 1$. A Bayes factor $\mathrm{BF}(t)$ based on a test statistic t is a so-called test-based Bayes factor (Johnson 2005) and often constitutes the most convenient way to transform a P-value to a Bayes factor. However, a test-based Bayes factor may not be equal to a P-based Bayes factor if the transformation from t to p is not one-to-one. Then the P-based Bayes factor (4) should be preferred, since it is directly based on the P-value, the quantity of interest.

1.2. Minimum Bayes factors

The distribution $f(p \,|\, H_1)$ in (4) may depend on unknown parameters η, say, and the maximum likelihood estimate $\hat{\eta}_{\mathrm{ML}}$ of η for the observed P-value p can then be used to determine the minimum P-based Bayes factor

$$\mathrm{minBF}(p) = \frac{f(p \,|\, H_0)}{\max_\eta f(p \,|\, \eta, H_1)} = \frac{f(p \,|\, H_0)}{f(p \,|\, \hat{\eta}_{\mathrm{ML}}, H_1)}. \tag{6}$$

Minimum Bayes factor: the smallest possible Bayes factor within a pre-specified class of prior distributions over alternative hypotheses.

If the transformation from t to p is one-to-one, then also the minimum test-based Bayes factor based on (5) will be the same as the minimum P-based Bayes factor (6), if $f_t(t(p) \,|\, \eta, H_1)$ can be derived from $f_p(p \,|\, \eta, H_1)$ with a change-of-variables. In principle, minimum Bayes factors can also be considered for data-based Bayes factors but the computation may be cumbersome if the distribution $f(y \,|\, H_1)$ depends on many unknown parameters.

The minimum Bayes factor is the smallest possible Bayes factor that can be obtained for a P-value p in a certain class of distributions considered under the alternative. As such it provides an objective lower bound on the Bayes factor, an example of an objective Bayes procedure (Berger 2006). Note that minimum Bayes factors have the same asymmetry as P-values as they can be used only to assess the (maximal) evidence against H_0, not for H_0. Examples of P-based minimum Bayes factors will be given in Section 2.3. Incidentally, the corresponding maximum Bayes factor does usually not exist since the marginal likelihood under the alternative does not have a strictly positive minimum for continuous distributions. This will be illustrated in the example described in Section 1.3.1 and Figure 1.

1.3. Examples

We now describe two clinical applications where a Bayesian calibration of P-values is of interest. The first example describes a well-designed confirmatory study, where a single P-value is available for the primary outcome of interest. In the second example many exploratory P-values are available from a logistic regression analysis with many potential predictors. Exploratory P-values are to be understood as summary statistics of the data only and should not be used for decision-making, but they can be used for generating hypotheses. The distinction between confirmatory and exploratory P-values is important (Berry 2016; Matthews et al. 2017) and requires different methods for a Bayesian calibration via minimum Bayes factors. We will argue that simple alternatives are suitable for confirmatory P-values, whereas local alternatives should be used for exploratory P-values.

1.3.1. Confirmatory P-values. Imagine a randomized controlled clinical trial, designed to detect a pre-specified clinically relevant difference with 80% power ($\beta = 0.2$) at the usual two-sided 5% significance level ($\alpha = 0.05$). A two-sided P-value $p = 0.01$ has been reported for the null hypothesis H_0 of no difference between the two treatments. The Principal Investigator (PI) of the trial knows that the P-value is only an indirect measure of the evidence against H_0 and has read a lot of the recent literature on misinterpretations and problems with P-values. He therefore asks the trial statistician to compute a Bayes factor as a direct measure of the evidence against the null. The statistician has calculated p based on a test statistic t which follows - for sufficiently large sample size - a standard normal distribution if H_0 is true. He has also derived the distribution of t under the assumption of the alternative H_1: $t \sim \mathrm{N}(\mu, 1)$ (Matthews 2006, Section 3.3) with

Folded normal distribution: the distribution of the absolute value of a normally distributed random variable.

$$\mu = \Phi^{-1}(1 - \alpha/2) + \Phi^{-1}(1 - \beta), \tag{7}$$

where $\Phi(.)$ denotes the standard normal cdf. However, the two-sided P-value $p = 2\,[1 - \Phi(|t|)]$ is not a one-to-one function of t, but is a one-to-one function of the absolute value of t. With a change-of-variables to the folded normal random variable $t^\star = |t|$ (see Appendix A.1 for its density function), the Bayes factor (5) then is

$$\mathrm{BF}(t^\star) = \frac{f(t^\star \,|\, H_0)}{f(t^\star \,|\, H_1)} = \frac{2\varphi(t^\star)}{\varphi(t^\star + \mu) + \varphi(t^\star - \mu)}, \tag{8}$$

where $\varphi(.)$ denotes the standard normal pdf and

$$t^\star = t^\star(p) = \Phi^{-1}(1 - p/2). \tag{9}$$

The trial statistician obtains $\mu = 2.80$ from equation (7) with $\alpha = 0.05$ and $\beta = 0.2$, $t^\star = 2.58$ from equation (9) with $p = 0.01$ and finally $\mathrm{BF}(p = 0.01) = 1/13$ (0.0744) from (8). He concludes that there is substantial evidence against H_0 since the probability of no effect has decreased from 50% (his prior guess) to $0.0744/(1 + 0.0744) = 6.9\%$.

However, the statistician is well aware that the assumptions underlying the sample size calculations may not be true. In particular, the power of the study may have been different from the assumed $1 - \beta$ value, if the true treatment effect is different from the pre-specified effect. He therefore minimizes (8) numerically with respect to μ (for fixed $t^\star = 2.58$) and obtains a lower bound $\mathrm{minBF}(p)$ on the Bayes factor, which turns out to be $\mathrm{minBF}(p) = 1/14$ (0.0725) with corresponding lower bound of 6.8% for $\mathsf{Pr}(H_0 \,|\, p)$.

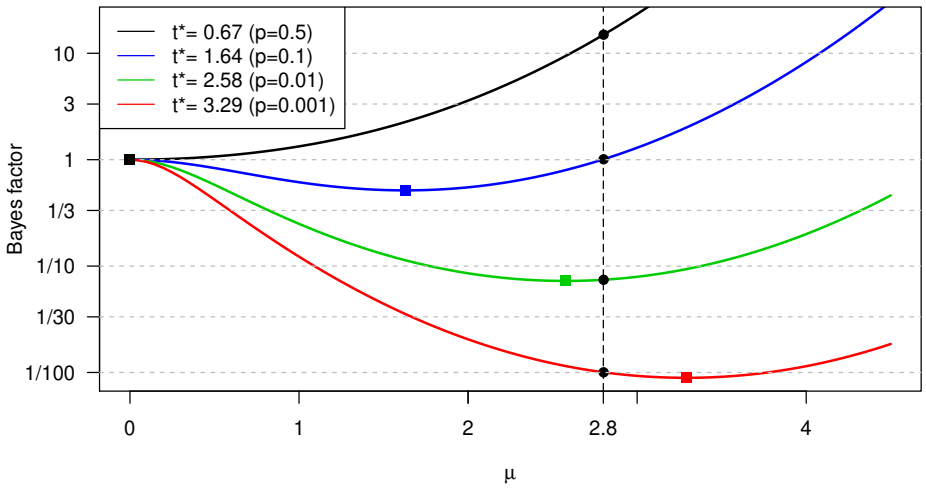

Figure 1

The Bayes factor (8) as a function of μ for $t^\star = 0.67$, 1.64, 2.58 and 3.29 ($p = 0.5$, 0.1, 0.01 and 0.001). The minima at $\hat{\mu}_{\text{ML}} = 0$, 1.63, 2.58 and 3.29, the values of μ that maximize $f(t^\star \mid H_1)$, are marked with squares and correspond to minimum Bayes factors of 1, 1/1.9, 1/14, and 1/112. Based on the sample size calculations we have $\mu = 2.80$ (the dashed black line) with Bayes factors of 15, 1, 1/13 and 1/100 (the black dots).

The step from the Bayes factor (8) to the minimum Bayes factor

$$\text{minBF}(t^\star) = \min_{\mu} \{\text{BF}(t^\star)\} \tag{10}$$

can be done for any value of $t^\star$, hence any P-value, as illustrated in Figure 1. Note that for $\mu = 0$ we have $\text{BF}(t^\star) = 1$ for any value of $t^\star$. The other extreme is $\mu \to \infty$, where $\text{BF}(t^\star) \to \infty$, again for any $t^\star$, so this Bayes factor has no maximum. Between these two extremes, there is a minimum of $\text{BF}(t^\star)$. For moderately large $t^\star$, say $t^\star \geq 1.64$, the minimum is near $\mu = t^\star$ (compare with Figure 1). If $t^\star \leq 1$, *i. e.* $p \geq 0.32$, the minimum is at $\mu = 0$ and $\text{minBF}(t^\star) = 1$ (Berger & Sellke 1987, Section 3.3).

Similar results can be obtained if the P-value $p = 0.01$ had been *one-sided* for the alternative H_1: $\mu > 0$. To see this, note that the Bayes factor now is

$$\text{BF}(t) = \frac{f(t \mid H_0)}{f(t \mid H_1)} = \frac{\varphi(t)}{\varphi(t-\mu)}, \tag{11}$$

where $t = \Phi^{-1}(1-p) = 2.33$. This leads to $\text{BF}(t) = 1/13$ (0.0748) and $\text{minBF}(t) = 1/15$ (0.0668).

So far, so good. But a colleague of the trial statistician notes that simpler procedures to compute a minimum Bayes factor based on a P-value have been proposed in the literature. Specifically, he mentions the "-e p log(p)" calibration (Sellke et al. 2001), which has been

reported to provide the lowest Bayes factor in favor of H_1 "under reasonable assumptions" (Bayarri et al. 2016). But surprisingly, for $p = 0.01$ this calibration gives a considerably larger minimum Bayes factor of 1/8 (0.125) than the Bayes factor 1/13 he has obtained. The assumptions underlying the sample size calculations have been thoroughly prepared and have been considered to be realistic by the PI and the ethics committee approving the trial protocol, so how can a lower bound on all reasonable Bayes factors be larger than his Bayes factor?

In fact, the "-e p log(p)" calibration closely agrees with the lower bound for local alternatives, but not for simple alternatives. The colleague thus points him to another calibration advocated by Goodman (1999b), who has proposed the lower bound $\exp(-t^2/2)$ for the Bayes factor where $|t| = t^\star$ as in (9). This bound turns out to be 1/28 for $p = 0.01$, so half as large as the lower bound 1/14 he has obtained. This seems overly conservative to the trial statistician, and he is now completely confused and unsure what Bayes factor he should report to the PI. We will see in Section 2.1 that the Goodman (1999b) bound, just as the minimum Bayes factor (10), is based on a simple alternative but incorporates additional knowledge on the direction of the effect.

1.3.2. Exploratory P-values. Many statistical procedures produce not only one, but a large number of P-values. For example, multiple regression, often used to develop clinical prediction models, gives a P-value for each potential predictor. For illustration we consider a publicly available subgroup of the GUSTO-I trial with $n = 2188$ patients (Steyerberg 2009). In order to develop a prediction model for the binary endpoint 30-day survival after acute myocardial infarction, we focus on the assessment of the effects of 17 covariates listed in Held et al. (2015, Table 1) using a logistic regression analysis. Note that 2 explanatory variables are categorical factor variables with 3 and 4 levels, respectively.

A first step to assess the importance of each of the predictors is to report 17 exploratory P-values in a standard regression table of the full model with all covariates. We will describe in Section 4.2 how such a table can be accompanied with the corresponding minimum Bayes factors. This analysis is exploratory in nature since the study was not powered for any of the potential predictors (treatment is not included), so we have a set-up where local alternatives should be used to calculate minimum Bayes factors.

1.4. Overview of Paper

In this paper we provide a comprehensive overview of different methods to transform P-values to minimum Bayes factors with an emphasis on two-sided P-based and test-based Bayes factors. We make the important distinction between simple and local alternatives, the latter class implying more restrictive assumptions, often leading to substantially larger minimum Bayes factors.

We start with a historical review in Section 2, where we describe the literature on data-based and P-based Bayes factors, as well as the more recent framework of test-based Bayes factors. The dependence of minimum Bayes factors on the sample size is described in Section 3. We will see that the maximal evidence of a P-value is inversely related to sample size. Test-based Bayes factors allow to investigate the dependence of the minimum Bayes factor on the dimension of θ (Section 4) and that minimum Bayes factor can be used to assess the combined evidence of multiple P-values (Section 4.3). We close with some discussion in Section 5 and some mathematical results are presented in the Appendix.

2. HISTORICAL REVIEW

Already Jeffreys (1961, Appendix B) studied the relationship between P-values and (approximate) Bayes factors for normally and binomially distributed observations (see also Berger & Sellke (1987) for more information on the normal case). One of the first papers with a systematic comparison of P-values and the corresponding minimum Bayes factors was already published in the 1960's (Edwards et al. 1963), considered as "still one of the best technical introductions to the Bayesian philosophy" (Spiegelhalter et al. 2004).

2.1. Simple Alternatives

Edwards et al. (1963, page 226) "examine one situation in which classical statistics prescribes a [...] t test", but in fact consider what is usually called a z-test for large samples (Bland 2015, Section 9.7). Specifically, the authors consider the problem of testing the point null hypothesis H_0: $\theta = \theta_0$ for a normally distributed observation $y \sim \mathrm{N}(\theta, \sigma^2)$ with mean θ and known variance σ^2. In practice the observation y will be a sufficient statistic for the parameter of interest, for example an average or a maximum likelihood estimate. The normality assumption underlies many statistical procedures found in medical journals (Goodman 1999b) and also in other areas of quantitative research. Let t denote the value of the t-statistic $t = (y - \theta_0)/\sigma$. Edwards et al. (1963) have observed that - for all possible prior densities $f(\theta \,|\, H_1)$ on θ under H_1, the Bayes factor for H_0 against H_1, $\mathrm{BF}(y) = f(y \,|\, H_0)/f(y \,|\, H_1)$ has the lower bound

$$\mathrm{minBF}(y) = \exp(-t^2/2), \tag{12}$$

the minimum Bayes factor in the class of all possible prior distributions for θ. They also note that the minimum is attained if "the density under the alternative hypothesis is concentrated at y, the place most favored by the data", *i. e.* for the simple alternative H_1: $\theta = \theta_1 = y$.

Note that the minimum Bayes factor in (12) is just a function of the test statistic t and that the corresponding test-based Bayes factor $\mathrm{BF}(t) = f(t \,|\, H_0)/f(t \,|\, H_1)$ based on $t \,|\, H_0 \sim \mathrm{N}(0, 1)$ and $t \,|\, H_1 \sim \mathrm{N}(\mu, 1)$ leads to the very same result, *i. e.* $\min_{\mu} \mathrm{BF}(t) = \mathrm{minBF}(y)$. It is often the case that a test-based Bayes factor is equal to the corresponding data-based Bayes factor, if the test statistic and prior distributions have been chosen carefully.

However, it is not clear how a P-value p should be transformed to the test statistic t. Edwards et al. (1963, page 228) restrict attention to one-sided P-values and use the corresponding one-sided t-value $t = \Phi^{-1}(1-p)$ in (12). This approach is based on the argument that "the alternative hypothesis has all its density on one side of the null hypothesis, [so] it is perhaps appropriate to compare the outcome of this procedure with the outcome of a one-tailed rather than a two-tailed classical test". We will see in Section 4.2.2 that the Edwards bound also provides a sharp lower bound on the Bayes factor under specific local alternatives for parameter vectors of any dimension, so-called g-priors.

In contrast, Goodman (1999b) has recommended to apply (12) to two-sided P-values to obtain the "smallest possible Bayes factor". The problem of this approach is that the test statistic t in (12) is not a one-to-one function of the two-sided P-value $p = 2[1 - \Phi(|t|)]$. Therefore the minimum Bayes factor (12) based on t is not the same as the minimum Bayes factor for the corresponding P-value p, since the former uses the additional information on the direction of the treatment effect, represented by the sign of t. The Goodman approach is therefore best seen as the Edwards bound applied to the corresponding one-sided P-value $p/2$, so that the information about the direction of the effect is included.

We have already described in Section 1.3.1 that the absolute test statistic $t^\star = |t|$ is a one-to-one function of p and the Bayes factor (8) based on $t^\star$ can be used to calculate the minimum Bayes factor (10) under a simple alternative, numerically minimizing (8) with respect to μ. This approach is equivalent to requiring the prior densities $f(\theta \,|\, H_1)$ to be symmetric (but possibly non-local) around θ_0, as described in Berger & Sellke (1987, Section 3.3). For two-sided P-values smaller than 0.1, so $t^\star > 1.64$, the minimum Bayes factor (10) can be well approximated as

$$\text{minBF}(t^\star) \approx \frac{2\varphi(t^\star)}{\varphi(2\,t^\star) + \varphi(0)} \approx 2\exp(-t^{\star\,2}/2) \tag{13}$$

(the exact multiplier on the right hand side of (13) is then between 1.99 and 2.0 rather than exactly 2). Comparing (13) with (12) we see that, for sufficiently small P-values, the Goodman (1999b) proposal is by a factor of 2 too small, and this is exactly what we have observed in the example considered in Section 1.3.1.

2.2. Local Alternatives

Another important case considered in Edwards et al. (1963) is a local normal prior for the mean θ of $y \sim \text{N}(\theta, \sigma^2)$, centered around the null value θ_0: $\theta \,|\, H_1 \sim \text{N}(\theta_0, \tau^2)$. This specification is appropriate for exploratory P-values from observational or hypotheses generating studies, where no specific alternative hypothesis has been specified *a priori*. It is shown that in this case the minimum Bayes factor (minimized with respect to τ^2, which yields $\tau^2 = \sigma^2 \max\{t^2 - 1, 0\}$) is

$$\text{minBF}(y) = \begin{cases} |t| \exp(-t^2/2)\sqrt{e} & \text{for } |t| > 1 \\ 1 & \text{otherwise,} \end{cases} \tag{14}$$

here $e \approx 2.72$ is Euler's number. Note that this bound also depends on the data only through the absolute value $t^\star = |t|$ of the t-statistic t. It is easy to show that we get the same result if we use the test-based Bayes factor based on $t^\star$, where $t^\star$ has a folded normal distribution with mean 0 and variance 1 under H_0 and it has a folded normal distribution with mean 0 and variance $1 + \tau^2/\sigma^2$ under H_1. Since the prior on θ under H_1 is centered around the null value θ_0, $t^\star$ has mean 0 under H_1 and it is easy to check that calculating the Bayes factor using t instead of $t^\star$, with $t \,|\, H_0 \sim \text{N}(0, 1)$ and $t \,|\, H_1 \sim \text{N}(0, 1 + \tau^2/\sigma^2)$, also leads to the same result. This is in contrast to the setting for bound (12), where the mean of $t^\star$ under H_1 is non-zero and the Bayes factors based on t and $t^\star$, respectively, differ. The "local normal alternatives" bound (14) is substantially larger than the Edwards bound (12), see Section 2.5.

We note that the more general class of all (possibly non-normal) local alternatives has been considered in Berger & Sellke (1987, Section 3.4). The resulting minimum Bayes factors are only slightly smaller than the ones obtained in the class of local normal priors. Local normal priors have the advantage that they can more easily be generalized to g-priors to investigate situations where the parameter of interest is a vector, see Sections 3.2 and 4.2.2.

2.3. P-based Bayes Factors

The minimum Bayes factors (12) and (14) depend on the value of the t-statistic t, so only indirectly on the P-value p. The commonly used "-e p log p" calibration, proposed in Vovk

(1993, Section 9) and Sellke et al. (2001) depends directly on the P-value p:

$$\text{minBF}(p) = \begin{cases} -e\,p\log p & \text{for } p < 1/e \\ 1 & \text{otherwise.} \end{cases} \tag{15}$$

A simple derivation of (15) assumes that under a point null hypothesis H_0, an exact P-value p is uniformly distributed on the unit interval. Under the alternative hypothesis, small P-values are expected, so the class of beta prior distributions $\text{Be}(\xi, 1)$ with monotonically decreasing density functions ($\xi \leq 1$) is considered.

The minimum Bayes factor (15) can then be derived as described in Sellke et al. (2001) and Held & Ott (2016, Appendix B), using the MLE $\hat{\xi}_{\text{ML}} = \min\{-1/\log(p), 1\}$. Sellke et al. (2001) also present an alternative derivation of (15), wherein one does not have to assume the beta class for the P-value under H_1. Held (2010) has noted that (15) can also be derived as a test-based Bayes factor under the g-prior if θ has dimension 2 and the sample size is large, see Section 4.2.2 for details. The calibration (15) is always smaller than the "local normal alternatives" bound (14) and approximately equal to the lower bound in the more general class of all local alternatives (Sellke et al. 2001, Section 3.2).

Prior sample size: the weight attached to a prior distribution, expressed as the equivalent sample size.

The beta distribution $\text{Be}(\xi, 1)$ has prior sample size $\xi + 1 \leq 2$, so is always quite uninformative. Therefore $f(p \,|\, H_1, \hat{\xi}_{\text{ML}})$ will be relatively flat and the minimum Bayes factor $\text{minBF}(p) = 1/f(p \,|\, H_1, \hat{\xi}_{\text{ML}})$ will be relatively large. However, this is not the only class of beta priors with monotonically decreasing density functions. An alternative, to our knowledge not yet discussed in the literature, is the class of beta distributions $\text{Be}(1, \kappa)$ with $\kappa \geq 1$. A beta distribution from this class has prior sample size $1 + \kappa \geq 2$, so the likelihood under the alternative can take larger values than for the above $\text{Be}(\xi, 1)$ prior. Calculus shows that in this setting the MLE of κ is $\hat{\kappa}_{\text{ML}} = \max\{-1/\log(1-p), 1\}$, leading to the minimum Bayes factor

$$\text{minBF}(p) = \begin{cases} -e\,(1-p)\log(1-p) & \text{for } p < 1 - 1/e \\ 1 & \text{otherwise.} \end{cases} \tag{16}$$

This is similar to the "- e p log p" calibration, but with p replaced by $q = 1 - p$, so we call this the "- e q log q" calibration. Note that for small enough p we can obtain the simple approximate formula $\text{minBF}(p) \approx e\,p$ based on the approximation $\log(1-p) \approx -p$.

It turns out that (16) is a much lower bound compared to all the other bounds proposed, for P-values less than 0.1 even smaller than the Goodman approach, see Section 2.5 for a comparison. This is due to a large (and unbounded) prior sample size for small p, in contrast to the prior sample size of the "- e p log p" calibration, which cannot be larger than 2. However, we will see in Section 3.2 that (16) provides a sharp lower bound on Bayes factors based on g-priors of any dimension d, even if the sample size is very small. For reasonably large sample sizes, however, the "- e q log q" calibration will be too conservative.

2.4. Test-based Bayes Factors

A drawback of data-based Bayes factors is that their values often depend critically on prior distributions that are assigned to unknown parameters under the null hypothesis and the alternative (Johnson 2005). Furthermore, computation of these Bayes factors may be involved as multi-dimensional integrals may need to be evaluated. In a landmark paper, Johnson (2005) proposes Bayes factors based on test statistics instead of the original data

to facilitate the use of Bayes factors. To obtain these Bayes factors, he considers the sampling distribution of the test statistic under the null and the alternative hypothesis. Usually, the distribution under the null does not depend on unknown model parameters and the distribution under the alternative can be parametrized in a parsimonious way - often as a "non-central" version of the distribution under the null hypothesis with only one additional non-centrality parameter, see Section 4.2 for an example. Thus, his approach eliminates the need to specify prior distributions for all unknown model parameters under each hypothesis and thus much of the subjectivity associated with Bayes factors. For several commonly used test statistics, he obtains a simple closed-form expression for the test-based Bayes factor assuming a computationally convenient prior for the non-centrality parameter. These results significantly simplify computation of Bayes factors. He considers χ^2-, F-, t- and z-test statistics in his 2005 paper, and extends the approach to likelihood ratio test (deviance) statistics in Johnson (2008), which allows for application of the methodology to generalized linear models (GLMs).

We will describe test-based Bayes factors based on the F-statistic in Section 3 and test-based Bayes factors based on the deviance in Section 4.2. A Bayesian model selection algorithm using test-based Bayes factors for linear models and GLMs is proposed in Hu & Johnson (2009). Held et al. (2015) show that Bayes factors based on the deviance statistic approximate data-based Bayes factors in GLMs and relate minimum test-based Bayes factors to minimum Bayes factors from the literature. There is also literature on Bayes factors based on nonparametric test statistics (Yuan & Johnson 2008). Bayes factors based on the deviance are applied to the Cox proportional hazard model in Held et al. (2016).

2.5. A Comparison

Edwards et al. (1963) compare P-values of 0.05, 0.01, and 0.001 with a selection of minimum Bayes factors. Table 3 provides a similar list with the minimum Bayes factors discussed so far, using the additional P-value $p = 0.005$ - as recently proposed by Benjamin et al. (2017) as a new threshold for statistical significance - and our preferred formatting of Bayes factors as ratios. First, note that all the minimum Bayes factors are substantially larger than the corresponding P-values and that the simple alternative minimum Bayes factor is always twice as large as the Goodman lower bound. Furthermore, the Edwards bound is close, but not equal to the simple alternative minimum Bayes factor. Also observe that the Goodman minimum Bayes factor for two-sided $p = 0.01$ is the same as the Edwards minimum Bayes factor for one-sided $p = 0.005$. The "-e p log p" bound is close to the "local normal alternatives" minimum Bayes factor. The "-e q log q" bound is smaller than all the other minimum Bayes factors, even smaller than the Goodman bound.

3. SAMPLE-SIZE ADJUSTED BAYES FACTORS

It is well-known that data-based Bayes factors depend on the sample size. By using Bayes factor methodology, several researchers have shown that the evidence of a P-value also depends on the underlying sample size (Jeffreys 1961; Royall 1986; Spiegelhalter et al. 2004; Wagenmakers 2007). In contrast, the P-based calibrations in Section 2.3 transform a given P-value to the same minimum Bayes factor no matter what the underlying sample size is. The same is true for the calibrations introduced in Sections 2.1 and 2.2 if the transforma-

P-value		0.05	0.01	0.005	0.001
Minimum Bayes factor	Formula				
Local normal alternatives	(14)	1/2.1	1/6.5	1/11	1/41
- e p log p	(15)	1/2.5	1/8	1/14	1/53
Simple alternative	(10)	1/3.4	1/14	1/26	1/112
Edwards	(12), one-sided	1/3.9	1/15	1/28	1/118
Goodman	(12), two-sided	1/6.8	1/28	1/51	1/224
- e q log q	(16)	1/7.5	1/37	1/74	1/368

Table 3 Comparison of P-values and various minimum Bayes factors. Table inspired by Table 4 in Edwards et al. (1963)

tion from the P-value to the test statistic is based on the quantiles of the (folded) normal distribution as described in those sections. However, the (folded) normal distribution is often only the asymptotic distribution of the test statistic. For small samples, such approximations should be avoided and the underlying sample size should be taken into account when transforming the P-value to the test-statistic and then to the minimum Bayes factor. Held & Ott (2016) have proposed such sample-size adjusted minimum Bayes factors for two-sided P-values from the t-test and F-test.

We will now describe the dependence of the minimum Bayes factors on sample size in several settings. In Section 3.1, we consider the t-test and the F-test in the linear model under simple alternatives. In Section 3.2, we study a class of local alternatives in the linear model, so-called g-priors, as in Held & Ott (2016).

3.1. Simple Alternatives

Let us revisit the motivating example from Section 1.3.1, which was based on a normal test statistic where a Bayes factor of $\text{BF}(p = 0.01) = 1/13$ (0.0744) with corresponding lower bound of 1/14 (0.0725) was obtained. A normal assumption is appropriate for large sample sizes, but suppose now that the sample size n of the study was fairly small, with only 10 patients in each group, so $n = 20$. Assume that the P-value $p = 0.01$ was obtained from the corresponding two-sample t-test with $n - 2$ degrees of freedom. The Bayes factor then has the form

$$\text{BF}(t^\star) = \frac{2f_{t(n-2)}(t^\star)}{f_{t(n-2)}(t^\star + \mu) + f_{t(n-2)}(t^\star - \mu)}, \tag{17}$$

where $f_{t(d)}(.)$ denotes the pdf of a standard t distribution with d degrees of freedom and $t^\star = t^\star(p)$ is now the $(1 - p/2)$-quantile of the standard t distribution with $n - 2$ degrees of freedom. Note that μ is computed as in (7), but with the standard t replacing the standard normal pdf at both occurrences. As in Section 1.3.1, we can minimize (17) with respect to μ to obtain the corresponding minimum Bayes factor

$$\text{minBF}(t^\star) = \min_{\mu} \{\text{BF}(t^\star)\} \tag{18}$$

under a simple alternative. The resulting Bayes factor (17), with $\mu = 2.96$, $n = 20$ and $t^\star(p = 0.01) = 2.88$, is $\text{BF}(t^\star = 2.88) = 1/18$ (0.0550), so somewhat smaller than for large sample size with a lower bound of $\text{minBF}(t^\star = 2.88) = 1/18$ (0.0548). This suggests that P-values obtained from small studies may carry more (maximal) evidence against the

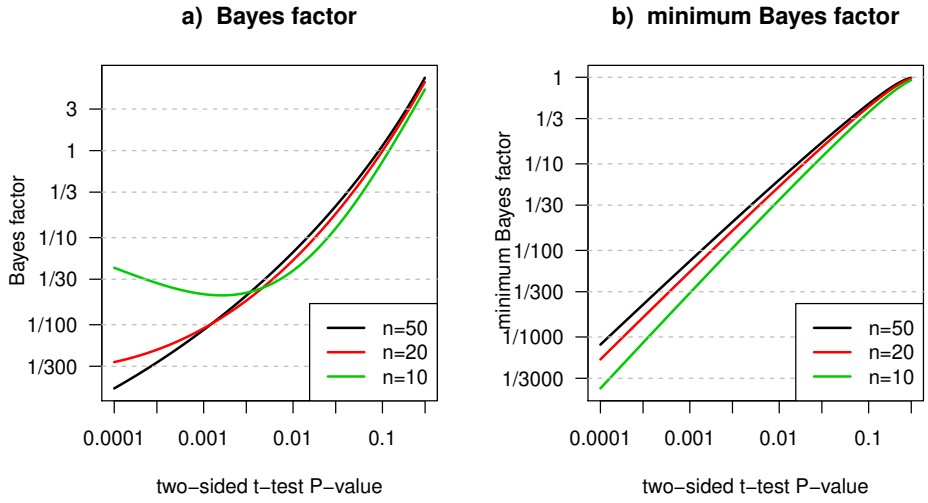

Figure 2

Bayes factors (a) with $\mu = 2.96$ and minimum Bayes factors (b) for a simple alternative as a function of the P-value from a t-test. Shown are the results for different sample sizes n.

null hypothesis than the very same P-values from larger studies, but this is only true for minimum Bayes factors, not for Bayes factors.

Indeed, Figure 2a illustrates that the Bayes factor (17) with fixed $\mu = 2.96$ of a small study can be larger than the Bayes factor of a larger study for small enough P-values. Spiegelhalter et al. (2004, Section 4.4.3) also observe a similar non-monotonic relationship of Bayes factors and sample size n for fixed P-values assuming a random sample of size n from a normal distribution and a local (normal) prior on the mean. In contrast, the minimum Bayes factor shown in Figure 2b decreases monotonically with decreasing sample size for any P-value.

If the P-value $p = 0.01$ in the motivating example had been *one-sided* for the alternative H_1: $\mu > 0$, then a sample-size adjusted modification of the large-sample Bayes factor (11) would be $\text{BF}(t) = f_{t(n-2)}(t)/f_{t(n-2)}(t-\mu)$, where the t-value t is now the $(1-p)$-quantile of the standard t distribution with $n-2$ degrees of freedom. For $\mu = 2.96$, $n = 20$ and $t(p = 0.01) = 2.55$, this Bayes factor is $\text{BF}(t) = 1/17$ (0.0581), which is similar to the value of the Bayes factor (17) for the two-sided P-value obtained above. The corresponding minimum Bayes factor

$$\text{minBF}(t) = \min_{\mu>0} \frac{f_{t(n-2)}(t)}{f_{t(n-2)}(t-\mu)} = \left(1 + \frac{t^2}{n-2}\right)^{-(n-1)/2} \tag{19}$$

turns out to be $\text{minBF}(t) = 1/19$ (0.0532), so only slightly smaller than the minimum Bayes factor (18) for the two-sided case. It is true in general that for the same large enough P-value (the threshold depending on the sample size n), the minimum Bayes factor (19) is

smaller than (18). For smaller P-values, these two minimum Bayes factors are very similar. Furthermore, the minimum Bayes factor (19) also decreases with decreasing sample size n for a fixed one-sided P-value from the t-test, so we observe the same dependence on n as for the two-sided t-test P-values.

The Bayes factor (17) is actually a special case of a Bayes factor for the linear model. It could alternatively be derived based on the F-statistic $f = (t^\star)^2$ instead of $t^\star$, where we have $f \sim F(1, n-2)$ under the null hypothesis. In the following, we will derive the Bayes factor for the F-test of overall significance in the standard linear regression model with intercept α,

$$\boldsymbol{y} = \alpha \mathbf{1} + \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{20}$$

where the response vector $\boldsymbol{y}$ is of length n, the regression coefficient vector $\boldsymbol{\theta}$ is of dimension $d \leq n-2$, the design matrix $\boldsymbol{X}$ has dimension $n \times d$ and the errors in $\boldsymbol{\epsilon}$ are assumed to be independent and normally distributed with zero mean and unknown residual variance σ^2. The F-statistic is then given as

$$f = \frac{R^2/d}{(1-R^2)/(n-d-1)}, \tag{21}$$

where R^2 is the usual coefficient of determination, the proportion of the variance in the response variable $\boldsymbol{y}$ that can be explained from the explanatory variables $\boldsymbol{X}$.

Under the null hypothesis H_0: $\boldsymbol{\theta} = \mathbf{0}$, the F-statistic (21) has a central F distribution with d and $n-d-1$ degrees of freedom, which is used to calculate the associated P-value, the upper tail probability at the observed F-value. Under the alternative H_1: $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, f has a non-central F distribution with d and $n-d-1$ degrees of freedom and non-centrality parameter λ. The resulting Bayes factor $\mathrm{BF}(f)$ can then be minimized by numerically optimizing the non-centrality parameter λ under the alternative to obtain the minimum Bayes factor over all possible simple alternatives. For $d = 1$, this approach corresponds to optimizing μ as in equation (18).

3.2. Local g-Priors

We now outline the derivation of a minimum test-based Bayes factor based on the F-statistic and local g-priors, as given in Johnson (2005). Suppose now we want to test the above null hypothesis H_0: $\boldsymbol{\theta} = \mathbf{0}$ against the composite alternative H_1: $\boldsymbol{\theta} \neq \mathbf{0}$. It is typically easier to assign a prior to the vector of regression coefficients $\boldsymbol{\theta}$ under H_1 than to the non-centrality parameter λ, the prior on $\boldsymbol{\theta}$ we will then imply a specific prior on λ. In the absence of substantive prior information, it is common to assign the g-prior (Zellner 1986)

$$\boldsymbol{\theta} \,|\, \sigma^2, H_1 \sim \mathrm{N}(\mathbf{0}, g\,\sigma^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \tag{22}$$

for fixed $g > 0$ to $\boldsymbol{\theta}$, which is invariant with respect to location-scale transformations of the covariates (Bayarri et al. 2012). Note that the g-prior is a local prior for H_0: $\boldsymbol{\theta} = \mathbf{0}$ with covariance matrix proportional to the inverse Fisher information matrix $\sigma^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ of the regression coefficients $\boldsymbol{\theta}$. It reduces to the normal prior described in Section 2.2 (for $\theta_0 = 0$) if θ is a scalar and σ^2 is known.

g-prior: a normal prior distribution with mean zero and covariance matrix proportional to the inverse Fisher information matrix of the regression coefficients.

For the following, no additional prior distribution on σ^2 is required since the prior distribution on λ implied by the g-prior (22) does not depend on σ^2 (we have $\lambda/g \sim \chi^2(d)$). By integrating out λ, one deduces that, under H_1, $f/(1+g)$ has a central F distribution

with d and $n-d-1$ degrees of freedom. The corresponding test-based Bayes factor turns out to be

$$\text{BF}(f) = (g+1)^{-(n-d-1)/2} \left\{ 1 + g \left[1 - \frac{f}{f + (n-d-1)/d} \right] \right\}^{(n-1)/2}. \tag{23}$$

Interestingly, the test-based Bayes factor (23) is equal to the data-based Bayes factor $\text{BF}(\boldsymbol{y})$ for the linear model (20) obtained under the g-prior (22) on $\boldsymbol{\theta} \,|\, \sigma^2$ combined with a reference prior $f(\alpha, \sigma^2) \propto \sigma^{-2}$ for the intercept α and the residual variance σ^2, as given in Liang et al. (2008). In particular, the data-based Bayes factor $\text{BF}(\boldsymbol{y})$ depends on the data only through the F-statistic (21), the sample size n and the dimension d of $\boldsymbol{\theta}$. The Bayes factor (23) can actually be derived under more general assumptions, where the null hypothesis is a linear constraint on the parameter vector $\boldsymbol{\theta}$, for example the null hypothesis that a single component of $\boldsymbol{\theta}$ is zero (Johnson 2005).

Note that $\text{BF}(f) = 1$ for $g = 0$ and $\text{BF}(f) \to \infty$ for $g \to \infty$. The first result is obvious as the Bayes factor then compares two identical models. The second result is related to the Jeffreys-Lindley paradox (Lindley 1957; Jeffreys 1961), which states that for large prior variances the Bayes factor always prefers the simpler model, no matter what the data are. In between there is a unique minimum of (23) for $\hat{g}_{\text{ML}} = \max\{f-1, 0\}$. By inserting the MLE $\hat{g}_{\text{ML}}$ into the Bayes factor (23), we obtain the minimum Bayes factor

$$\text{minBF}(f) = \min_g \text{BF}(f) = \begin{cases} \left[\frac{1+(n-d-1)/d}{f+(n-d-1)/d} \right]^{(n-1)/2} f^{d/2} & \text{for } f > 1 \\ 1 & \text{otherwise.} \end{cases} \tag{24}$$

Note that this formula only depends on f, n and d, so it provides a convenient way to transform an F-statistic (or the corresponding P-value) to a lower bound on the Bayes factor.

Held & Ott (2016) have studied the relationship between a P-value from the F-test and the corresponding minimum Bayes factor (24). Their main findings have been:

1. For fixed p and fixed dimension d, the minimum Bayes factor (24) decreases with decreasing sample size n.
2. For fixed p and fixed sample size n, the minimum Bayes factor (24) decreases with increasing dimension d.

Figure 3 compares the minimum Bayes factor (24) based on the local g-prior for $d = 1$ with the minimum Bayes factor (18) based on simple alternatives for fixed P-value and varying sample size n. We see the same pattern in both cases, with increasing minimum Bayes factors for increasing sample size.

In Figure 4 we study the dependence of the minimum Bayes (24) factor on the P-value for $d = 3$ and $d = 4$ and different sample sizes n, see Held & Ott (2016) for the corresponding plots for $d = 1$ and $d = 2$. We have added the "-e p log p" bound (15) as a blue line, which is always larger than the sample-size adjusted minimum Bayes factors. We have also added the "-e q log q" bound (16) as a red line, which is always below the sample-size adjusted minimum Bayes factors.

As a consequence of the Held & Ott (2016) results, the minimum Bayes factor (24) is largest for $d = 1$ and large n. As we will see in Section 4, the value of the minimum Bayes factor in fact converges for $n \to \infty$ to the "local normal alternatives" bound (14) with $t^\star = \sqrt{f}$. On the other hand, the minimum Bayes factor (24) is smallest for large

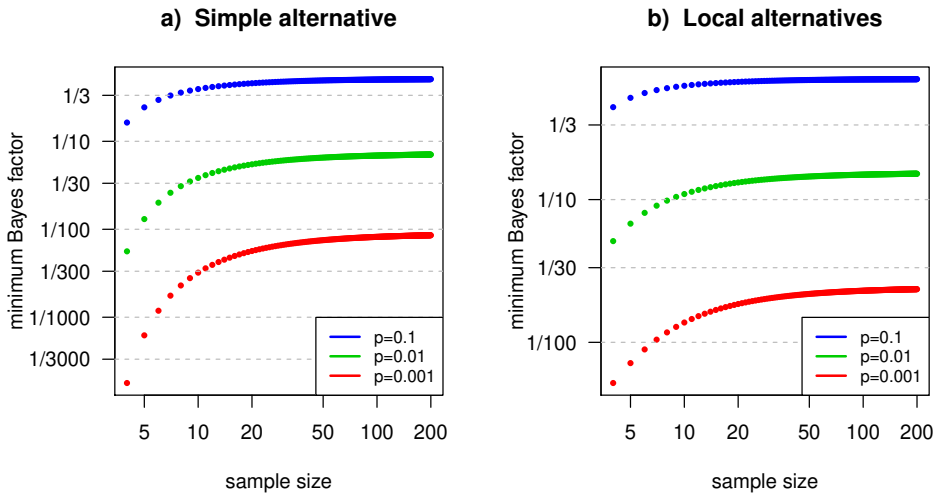

Figure 3

The dependence of the minimum Bayes factors (18) (a) and (24) with $d = 1$ (b) on sample size for fixed P-value.

d and small n. Standard regularity conditions in the linear model require $n \geq d + 2$ and we will now consider the case $n = d + 3$, where $f \mid H_0 \sim F(d, 2)$. The quantile function of $f \mid H_0$ is then available in closed form (see equation (35) in Appendix A.2.1), and there is a closed-form expression for the minimum Bayes factor (24) as a function of the P-value p:

$$\text{minBF}(p) = (d+2)\,\frac{1-(1-p)^{2/d}}{2}\,\left(\frac{d+2}{d}\right)^{d/2}(1-p), \tag{25}$$

as derived in Appendix A.2.1. We show in Appendix A.2.1 that the limit of (25) for $d \to \infty$ is the "- e q log q" calibration (16). Since the convergence is from above, the "- e q log q" calibration (16) is a universal lower bound on sample-size adjusted minimum Bayes factors based on local g-priors, if we exclude the very extreme case $n = d + 2$, where the minimum Bayes factor can be even smaller than (16).

Another interesting special case of the minimum Bayes factor (24) based on the F-statistic can be obtained for $d = 2$. In this case, there is a closed-form relationship between the P-value from the F-test and the F-statistic (Held & Ott 2016, equation (11)), so (24) can be rewritten as a P-based Bayes factor:

$$\text{minBF}(p) = \frac{1}{2}\left[\frac{(n-1)^{(n-1)}}{(n-3)^{(n-3)}}\right]^{1/2}\left[1-p^{2/(n-3)}\right]\,p \tag{26}$$

$$\approx \frac{e}{2}(n-2)\left[1-p^{2/(n-3)}\right]\,p \tag{27}$$

for $p < (\frac{n-1}{n-3})^{-(n-3)/2}$, otherwise $\text{minBF}(p) = 1$. Held & Ott (2016) show that for fixed n, (26) is always smaller than the "-e p log p" calibration (15) and that (27) converges from

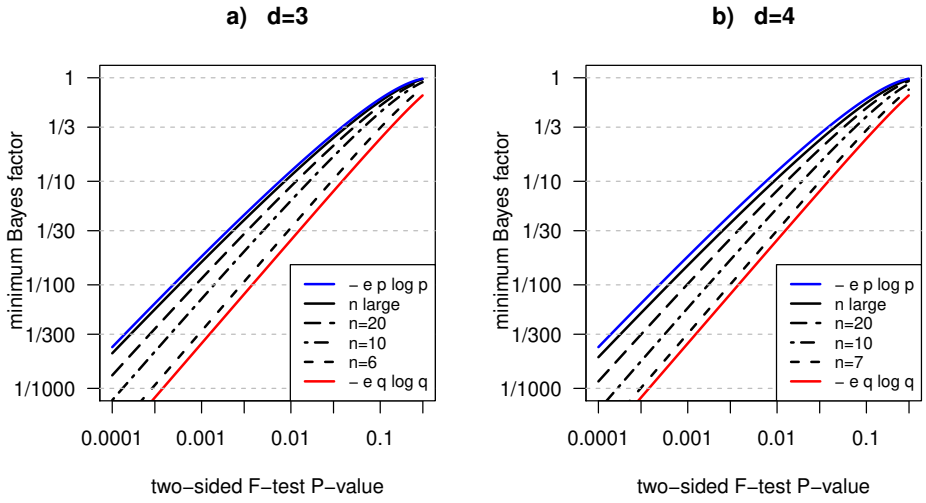

Figure 4

Minimum Bayes factors (24) based on local g-priors as a function of the P-value from an F-test. Shown are the bounds for $d = 3$ (a) and $d = 4$ (b) for sample size $n = d + 3, 10, 20$ and for n large. The blue line is the "-e p log p" calibration (15). The red line is the "-e q log q" calibration (16).

below to (15) for $n \to \infty$. However, it has been argued that already the "- e p log(p)" calibration (15) provides a (lower) bound on the Bayes factor "under general assumptions" (Stephens & Balding 2009, page 684) and constitutes "a best-case scenario for the strength of the evidence in favor of H_1 that can arise from a given p-value" (Bayarri et al. 2016, page 91). This is not true for g-priors, as our analysis has shown. As illustrated in Figure 4, the minimum Bayes factor (24) is always smaller than (15) for any $d \geq 2$ and any finite sample size. Even for $d = 1$, the standard t-test setting, the calibration (15) can be larger than (24) if n is small. For example, for P-values not smaller than 10^{-4}, the sample size must be $n = 27$ or larger, such that (15) is a valid bound. For P-values not smaller than 10^{-6}, the sample size must be $n = 37$ or larger for (15) to be valid.

4. LARGE-SAMPLE BAYES FACTORS

We have seen in Section 3 that the (approximate) minimum Bayes factor (27) converges to the "-e p log p" calibration (15) for $n \to \infty$. We will now generalize that result by establishing convergence of the minimum Bayes factor (24) to a test-based Bayes factor based on the deviance for general d. We will also provide an alternative derivation of that test-based Bayes factor in the GLM framework and analyze its dependence on d.

4.1. Some Convergence Results

It is easy to see that the Bayes factor (17) converges to the Bayes factor (8) as the sample size n goes to infinity, since the absolute value $t^\star = |t|$ of the t-statistic converges to (9) and the t-density in (17) converges to a standard normal density as the degrees of freedom go to infinity. The Bayes factor (17) is a special case (for $d = 1$) of the Bayes factor based on the F-statistic under simple alternatives.

Next, we study the Bayes factor (23) based on the F-statistic, which was derived under the g-prior, as the sample size n goes to infinity. To do so, we assume a sequence of alternatives of the form H_1^n: $\boldsymbol{\theta} = \mathcal{O}(n^{-1/2})$ in the linear model (20), so the size of the true regression coefficients $\boldsymbol{\theta}$ gets smaller with increasing sample size n. This is the case of practical interest, because for larger $\boldsymbol{\theta}$ it would be trivial to differentiate between H_0: $\boldsymbol{\theta} = \mathbf{0}$ and H_1^n, and for smaller $\boldsymbol{\theta}$ it would be too difficult (Johnson 2005, page 691). Under such a sequence of alternatives, the coefficient of determination R^2 and the F-statistic f tend to zero as the sample size goes to infinity. In contrast, the deviance (or likelihood ratio test) statistic

$$z = 2 \log \left[\frac{\max_{\alpha, \boldsymbol{\theta}} f(\boldsymbol{y} \,|\, \alpha, \boldsymbol{\theta}, H_1)}{\max_{\alpha} f(\boldsymbol{y} \,|\, \alpha, H_0)} \right]$$

has a limiting distribution in this setting (Johnson 2008), so it is of order $\mathcal{O}(1)$. For fixed deviance z and fixed $g > 0$, we then obtain

$$\lim_{n \to \infty} \mathrm{BF}(f) = \mathrm{BF}(z), \tag{28}$$

where

$$\mathrm{BF}(z) = (g+1)^{d/2} \exp\left(-\frac{g}{g+1} \frac{z}{2} \right) \tag{29}$$

is a test-based Bayes factor based on the deviance z, see Appendix A.2.2 for the proof of this convergence result.

4.2. Generalized Linear Models

As mentioned in Section 2.4, test-based Bayes factors based on the deviance can be applied in a wider context, including generalized linear models. To keep notation simple, we consider a generalized linear model with linear predictor vector $\eta = \alpha \mathbf{1} + \boldsymbol{X}\boldsymbol{\theta}$ and test H_0: $\boldsymbol{\theta} = \mathbf{0}$ against the alternative H_1: $\boldsymbol{\theta} \neq \mathbf{0}$. However, the approach can easily be generalized to null hypotheses where only subvectors of $\boldsymbol{\theta}$ are fixed (Hu & Johnson 2009).

Under regularity conditions we have the well-known result that under H_0 the deviance z has an asymptotic chi-squared distribution $\chi^2(d)$ with d degrees of freedom, where $d = \dim(\boldsymbol{\theta})$ is the dimension of the parameter of interest. The deviance $z = z(p)$ is then a one-to-one function of the corresponding P-value $p = \Pr(\chi^2(d) \geq z)$.

To obtain the limiting distribution under the alternative H_1, we again consider alternatives of the form H_1^n: $\boldsymbol{\theta} = \mathcal{O}(n^{-1/2})$. Under such a sequence of alternatives and some regularity conditions, the distribution of the deviance converges to a non-central chi-squared distribution with d degrees of freedom and non-centrality parameter $\lambda = \mathcal{O}(1)$ (Davidson & Lever 1970; Held et al. 2015).

4.2.1. Simple alternatives. The test-based Bayes factor $\mathrm{BF}(z) = f(z \,|\, H_0)/f(z \,|\, H_1)$ compares the likelihood of z under the asymptotic central and non-central chi-squared distribu-

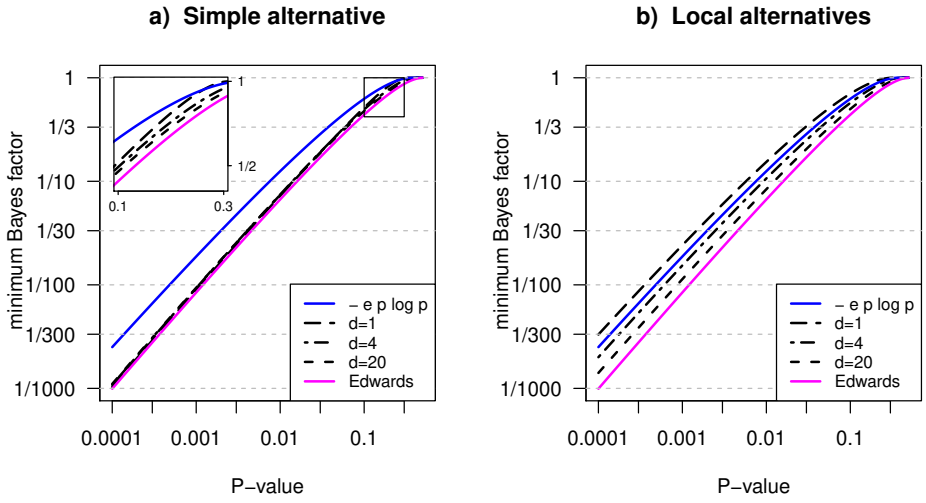

Figure 5

Minimum Bayes factors based on the deviance for different dimensions $d = \dim(\boldsymbol{\theta})$ under simple (a) and local (b) alternatives as a function of P-values. Under local alternatives based on the g-prior, the "-e p log p" calibration (15) (blue line) is obtained for dimension $d = 2$ and the Edwards minimum Bayes factor (12) (pink line) for $d \to \infty$.

tion. The corresponding minimum Bayes factor can be obtained numerically by maximizing the non-central chi-squared density of z under the alternative H_1 with respect to λ.

The minimum Bayes factors are very similar for different dimensions d, see Figure 5a. For larger P-values we see the expected ordering of the minimum Bayes factors with larger values for smaller d. For $p < 0.1$, the minimum Bayes factors are all below the "-e p log p" calibration, but only slightly larger than the Edwards bound.

4.2.2. Local alternatives. Expression (29) can also be derived directly as a test-based Bayes factor based on the deviance statistic z, as proposed in Johnson (2008). Assume the generalized g-prior $\boldsymbol{\theta} \,|\, H_1 \sim \mathrm{N}(\mathbf{0}, g\boldsymbol{I}^{-1}_{\boldsymbol{\theta},\boldsymbol{\theta}})$, where $g > 0$ and $\boldsymbol{I}_{\boldsymbol{\theta},\boldsymbol{\theta}}$ denotes the expected Fisher information matrix for $\boldsymbol{\theta}$. This prior is only used implicitly in the derivation and corresponds to a gamma prior $\lambda \sim \mathrm{G}(d/2, 1/[2g])$ (with mean $d \cdot g$) on the non-centrality parameter $\lambda = \boldsymbol{\theta}^\top \boldsymbol{I}_{\boldsymbol{\theta},\boldsymbol{\theta}}\, \boldsymbol{\theta}$ of the asymptotic non-central chi-squared distribution for the deviance z under the sequence of alternatives H_1^n. The implied approximate marginal distribution of z is then

$$z \,|\, H_1 \overset{a}{\sim} \mathrm{G}\left(d/2, 1/[2(g+1)]\right) \tag{30}$$

(Johnson 2008, theorem 2), which serves as marginal likelihood $f(z \,|\, H_1)$ under the alternative. Furthermore, we have the approximate likelihood $f(z \,|\, H_0)$ under the null hypothesis from $z \overset{a}{\sim} \mathrm{G}(d/2, 1/2) = \chi^2(d)$. With these prerequisites, we can derive the test-based Bayes factor $\mathrm{BF}(z) = f(z \,|\, H_0)/f(z \,|\, H_1)$ (29) of H_0 versus H_1 for fixed g. For example,

for $d = 1$ and large sample size n, formula (29) is equivalent to formula (7) proposed by Wakefield (2009) in the context of genome-wide association studies (since the deviance and the squared Wald statistic are asymptotically equivalent).

To determine the minimum Bayes factor, we maximize the marginal likelihood (30) under H_1 with respect to g and obtain the estimate

$$\hat{g}_{\text{ML}} = \max\{z/d - 1, 0\}. \tag{31}$$

Inserting (31) into (29) then gives (Johnson 2008; Held et al. 2015)[3]

$$\text{minBF}(z) \begin{cases} \left(\frac{z}{d}\right)^{d/2} \exp\left(-\frac{z-d}{2}\right) & \text{for } z > d \\ 1 & \text{otherwise.} \end{cases} \tag{32}$$

For any fixed P-value $p = \Pr(\chi^2(d) \geq z)$, the minimum Bayes factor (32) decreases monotonically as d increases (see Figure 5b). In some special cases, the minimum Bayes factor (32) corresponds to minimum Bayes factors from the literature introduced in Section 2, all shown in Figure 5b: For $d = 1$, $\text{minBF}(z)$ is equivalent to the "local normal alternatives" bound (14) and Held et al. (2015) show that for $d = 2$, where $z = -2\log(p)$, $\text{minBF}(z)$ reduces to the "- e p log p" calibration (15). Furthermore, as the dimension d tends to infinity, $\text{minBF}(z)$ tends to the Edwards minimum Bayes factor $\text{minBF}(t)$ (12) with one-sided t-value, see Appendix A.3.1 for the proof. The same dependence of the minimum Bayes factor on the dimension d has been reported in Sellke et al. (2001, Table 4) for a slightly different class of local priors.

We now return to the example described in Section 1.3.2. Application of Bayes factors based on the deviance to a logistic regression results in Table 4, where we list for each covariate the corresponding deviance z, the dimension d of the parameter of interest, the P-value p and the minimum Bayes factor (32). Note that the deviance is always calculated based on a comparison of the full model with the model where the covariate of interest has been removed. Note also that $d = 3$ for the factor variable "Killip class" with four levels, $d = 2$ for the factor variable "Smoking" with three levels (Never/Ex/Current) and $d = 1$ for the remaining variables.

There are six variables with no evidence and another five variables with only weak evidence for an association with the outcome. Three covariates show overwhelming evidence for an association ($\text{minBF} < 1/1000$), the remaining three covariates show moderate to substantial evidence with minimum Bayes factors between 1/4.9 and 1/19.

4.3. Combining Evidence

Suppose now that several two-sided P-values $p_1, \ldots, p_n$ are available from n independent studies, for example from different clinical trials to investigate the efficacy of the same treatment. How can we combine the statistical evidence available from those studies into one minimum Bayes factor? Deviance-based Bayes factors provide a convenient tool for doing so. One option would be to compute the minimum Bayes factor (32) based on the deviance $z_i = z(p_i)$ for each P-value p_i with associated dimension d_i, and then compute

[3]This is a correction of the formula given in Held et al. (2015) in the case $z \leq d$.

	Deviance z	Dimension d	P-value p	min Bayes factor
Gender	2.75	1	0.097	1/1.4
Age	75.72	1	< 0.0001	$< 1/1000$
Killip class	38.68	3	< 0.0001	$< 1/1000$
Diabetes	0.19	1	0.67	1
Hypotension	19.33	1	< 0.0001	$< 1/1000$
Tachycardia	9.12	1	0.003	1/19
Anterior infarct location	1.93	1	0.16	1/1.1
Previous myocardial infarction	5.97	1	0.015	1/4.9
Height	0.63	1	0.43	1
Weight	2.28	1	0.13	1/1.3
Hypertension history	1.93	1	0.16	1/1.1
Smoking history	0.66	2	0.72	1
Hypercholesterolaemia	0.51	1	0.48	1
Previous Angina Pectoris	0.68	1	0.41	1
Family history	0.45	1	0.50	1
ST elevation on ECG	8.72	1	0.003	1/16
Persistent chest pain	1.72	1	0.19	1/1.1

Table 4 Output from a logistic regression model to identify important predictors of 30-day survival in the GUSTO-I study. Shown is the deviance z with dimension d of the parameter of interest, the P-value p and the associated minimum Bayes factor (32) for local alternatives based on the g-prior

the overall minimum Bayes factor as the product

$$\prod_{i=1}^{n} \text{minBF}(z_i). \tag{33}$$

To see why we are taking the product of the minimum Bayes factors, note that the Bayes factor for the combined evidence equals the product of the Bayes factors $\text{BF}(z_i)$, $i = 1, \ldots, n$, for the single studies by sequential updating of Bayes factors (see Goodman (2016) for a practical example). We are interested in the minimum of this product Bayes factor, which has the product of the single minimum Bayes factors $\text{minBF}(z_i)$ as a lower bound.

A sharper bound can be obtained by an application of Fisher's method to combine P-values from independent studies (Fisher 1958). He suggested to compute $z_+ = \sum_{i=1}^{n} z_i$ where $z_i = -2\log(p_i)$ is in fact the deviance test statistic with $d = 2$. Fisher argued that, under H_0, each z_i follows a chi-squared distribution with 2 degrees of freedom, so z_+ is chi-squared with $d_+ = 2n$ degrees of freedom, which can be used to calculate a combined P-value $p_+ = \Pr(\chi^2(2n) \geq z_+)$. To calculate the associated minimum Bayes factor, we can therefore use (32) (with $d = 2n$) based on z_+. This approach gives a sharper (*i. e.* larger) bound than the product minBF with equality if all P-values are identical. The approach can in fact be applied for any dimensions d_i, $i = 1, \ldots, n$, then $d_+ = \sum_{i=1}^{n} d_i$ and the same inequality between the product minBF and the combined minBF still holds, see Appendix A.3.2 for the proof.

For illustration, consider the original example from Fisher (1958, § 22.1), where three tests of significance have yielded the P-values $p_1 = 0.145$, $p_2 = 0.263$, and $p_3 = 0.087$, and Fisher's method gives the combined P-value $p_+ = 0.076$. The product minimum Bayes factor (33) (with $d = 2$) is then 1/2.4 (0.42) whereas the minimum Bayes factor based on the combined P-value (with $d = 6$) is 1/2.2 (0.46), so slightly larger in accordance with the proof in Appendix A.3.2. If instead the P-values are based on dimension $d = 1$, then the combined P-value is $p_+ = 0.098$, the product minimum Bayes factor (33) is 1/1.9 (0.53)

whereas the minimum Bayes factor based on the combined P-value (now with $d = 3$) is 1/1.7 (0.58).

5. DISCUSSION AND OUTLOOK

The main findings of this review are summarized as summary points below. We close now with two extensions of the methodology described.

5.1. Sample-size Adjusted Bayes Factors in GLMs

For GLMs, marginal likelihoods under local priors on the vector of regression coefficients $\boldsymbol{\theta}$, such as generalized g-priors, are typically not available in closed form, so they need to be computed by numerical techniques (numerical integration or Monte Carlo methods). Bayes factors based on the deviance statistic are therefore especially appealing for GLMs as they significantly simplify computations. However, these Bayes factors are not adjusted for sample size.

An alternative approach, which allows for sample size adjustments and is also computationally efficient, is to derive approximate data-based Bayes factors in closed form by applying analytical approximations, so called integrated Laplace approximations (Wang & George 2007; Li & Clyde 2016). For example, by applying the Li & Clyde (2016) methodology, an approximate, sample-size adjusted minimum Bayes factor for 2×2 contingency tables can be obtained in closed form (Ott & Held 2017). By studying the relationship between this minimum Bayes factor and two-sided P-values from Fisher's exact test, Ott & Held (2017) conclude that the maximal evidence of these P-values is inversely related to sample size. This is the same qualitative relationship as in the linear model, see Section 3.2 and Figure 4.

5.2. Interval Null Hypotheses

One criticism of point null significance testing is that exact point null hypotheses rarely arise in practice. Instead, researchers often aim at testing if a parameter is close to the null value θ_0, which corresponds to an interval null hypothesis of the form H_0: $\theta \in (\theta_0 - b, \theta_0 + b)$ for some small b. However, Berger & Sellke (1987, page 114) argue that "for a large number of problems testing a point null hypothesis is a good *approximation* to the actual problem". They state that if b is small, the minimum Bayes factor for the interval null hypothesis H_0: $\theta \in (\theta_0 - b, \theta_0 + b)$ is essentially equivalent to the minimum Bayes factor for the corresponding point null hypothesis H_0: $\theta = \theta_0$ if the same class of alternatives is considered. A similar argument is provided by Johnson (2016).

SUMMARY POINTS

1. P-values are indirect measures of the evidence against a point null hypothesis H_0. Bayes factors provide a quantitative summary of the direct evidence against H_0.
2. P-values can be transformed to minimum Bayes factors. A minimum Bayes factor quantifies the maximal evidence of a P-value against a point null hypothesis within a certain class of alternative hypotheses.
3. The maximal evidence of a P-value depends on how the P-value has been calculated.

It generally decreases with increasing sample size, but increases with increasing dimension of the parameter of interest. These features should be taken into account when P-values are transformed to minimum Bayes factors in routine applications.

4. The maximal evidence of a P-value also depends on the underlying study design: It matters whether the P-value comes from a confirmatory study with a well-defined simple alternative, or from an exploratory analysis used to generate hypotheses, where local alternatives are more appropriate.
5. The commonly used "- e p log(p)" calibration represents a lower bound on the Bayes factor for local alternatives on scalar parameters ($d = 1$) in large samples, but not necessarily for small samples or for larger dimensions of the parameter of interest.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation [project #159715].

LITERATURE CITED

Bayarri MJ, Benjamin DJ, Berger JO, Sellke TM. 2016. Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *J. Math. Psychol.* 72:90–103

Bayarri MJ, Berger JO, Forte A, García-Donato G. 2012. Criteria for Bayesian model choice with application to variable selection. *Ann. Stat.* 40:1550–1577

Benjamin DJ, Berger JO, Johannesson M, Nosek, BA, Wagenmakers E-J et al. 2017. Redefine statistical significance. *Nature Human Behaviour* http://dx.doi.org/10.1038/s41562-017-0189-z

Berger JO. 2006. The case for objective Bayesian analysis. *Bayesian Anal.* 1:385–402

Berger JO, Delampady M. 1987. Testing precise hypotheses. *Stat. Sci.* 2:317–335

Derivation of minimum Bayes factors for different classes of alternatives, including symmetric and local alternatives

Berger JO, Sellke T. 1987. Testing a point null hypothesis: The irreconcilability of P values and evidence (with discussion). *J. Am. Stat. Assoc.* 82:112–139

Bernardo JM, Smith AFM. 2000. Bayesian Theory. Wiley Ser. Prob. Stat. Chichester: John Wiley & Sons

Berry DA. 2016. P-values are not what theyre cracked up to be. *Am. Stat.* 70. Online Discussion in Supplemental Material

Bland M. 2015. An Introduction to Medical Statistics. Oxford: Oxford University Press, 4th ed.

Casella G, Berger RL. 1987. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Am. Stat. Assoc.* 82:106–111

Cox DR. 2006. Principles of Statistical Inference. Cambridge: Cambridge University Press

Cox DR, Donnelly CA. 2011. Principles of Applied Statistics. Cambridge: Cambridge University Press

A celebrated introduction to the Bayesian paradigm includes a pioneering section on minimum Bayes factors

Davidson RR, Lever WE. 1970. The limiting distribution of the likelihood ratio statistic under a class of local alternatives. *Sankhya Ser. A* 32:209–224

Donahue RMJ. 1999. A note on information seldom reported via the P value. *Am. Stat.* 53:303–306

Edwards W, Lindman H, Savage LJ. 1963. Bayesian statistical inference for psychological research. *Psychol. Rev.* 70:193–242

Fisher RA. 1958. Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd, 13th ed.
Good IJ. 1950. Probability and the Weighing of Evidence. London: Griffin
Goodman SN. 1999a. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann. Intern. Med.* 130:995–1004
Goodman SN. 1999b. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.* 130:1005–1013
Goodman SN. 2005. P value. In *Encyclopedia of Biostatistics*. Chichester: Wiley, 2nd ed., 3921–3925
Goodman SN. 2008. A dirty dozen: Twelve p-value misconceptions. *Semin. Hematol.* 45:135–140
Goodman SN. 2016. Aligning statistical and scientific reasoning. *Science* 352:1180–1181
Greenland S, Poole C. 2013. Living with P values: Resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology* 24:62–68
Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, et al. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* 31:337–350
Held, L. (2010). A nomogram for P values. *BMC Med. Res. Methodol.* 10:21
Held L, Gravestock I, Sabanés Bové D. 2016. Objective Bayesian model selection for Cox regression. *Stat. Med.* 35:5376–5390
Held L, Ott M. 2016. How the maximal evidence of P-values against point null hypotheses depends on sample size. *Am. Stat.* 70:335–341
Held L, Sabanés Bové D, Gravestock I. 2015. Approximate Bayesian model selection with the deviance statistic. *Stat. Sci.* 30:242–257
Hu J, Johnson VE. 2009. Bayesian model selection using test statistics. *J. R. Stat. Soc. Ser. B* 71:143–158
Hung HMJ, O'Neill RT, Bauer P, Kohne K. 1997. The behavior of the P-value when the alternative hypothesis is true. *Biometrics* 53:11–22
Jeffreys H. 1961. Theory of Probability. Oxford: Oxford University Press, 3rd ed.
Johnson VE. 2005. Bayes factors based on test statistics. *J. R. Stat. Soc. Ser. B* 67:689–701
Johnson VE. 2008. Properties of Bayes factors based on test statistics. *Scand. J. Stat.* 35:354–368
Johnson VE. 2016. Comments on the "ASA Statement on Statistical Significance and P-values" and marginally significant P-values. *Am. Stat.* 70. Online Discussion in Supplemental Material
Johnson VE, Rossell D. 2010. On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B* 72:143–170
Kass R, Raftery A. 1995. Bayes factors. *J. Am. Stat. Assoc.* 90:773–795
Lee PM. 2004. Bayesian Statistics: An Introduction. London: Wiley, 3rd ed.
Li Y, Clyde MA. 2016. Mixtures of g-priors in generalized linear models. Tech. rep., Clemson/Duke University
Liang F, Paulo R, Molina G, Clyde MA, Berger JO. 2008. Mixtures of g priors for Bayesian variable selection. *J. Am. Stat. Assoc.* 103:410–423
Lindley DV. 1957. A statistical paradox. *Biometrika* 44:187–192
Marsman M, Wagenmakers EJ. 2017. Three insights from a Bayesian interpretation of the one-sided P value. *Educ. Psychol. Meas.* 77:529–539
Matthews JN. 2006. Introduction to Randomized Controlled Clinical Trials. Texts in Stat. Sci. Boca Raton: Chapman & Hall/CRC, 2nd ed.
Matthews R, Wasserstein R, Spiegelhalter D. 2017. The ASA's p-value statement, one year on. *Significance* 14:38–41
Ott M, Held L. 2017. Bayesian calibration of P-values from Fisher's exact test. Tech. rep., University of Zurich. Submitted
Ramsey F, Schafer D. 2002. The Statistical Sleuth: A Course in Methods of Data Analysis. Belmont, California: Duxbury Press, 2nd ed.
Royall RM. 1986. The effect of sample size on the meaning of significance tests. *Am. Stat.* 40:313–315

Two papers advocating minimum Bayes factors as an alternative to P-values in medical research

A sample-size adjusted calibration of P-values is proposed

Bayes factors based on test statistics are introduced

A comprehensive paper on the "$-ep\log p$" calibration gives different derivations

Sellke T, Bayarri MJ, Berger JO. 2001. Calibration of p values for testing precise null hypotheses. *Am. Stat* 55:62–71

Spiegelhalter DJ, Abrams KR, Myles JP. 2004. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. New York: Wiley

Stephens M, Balding DJ. 2009. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* 10:681–690

Steyerberg E. 2009. Clinical Prediction Models. Stat. for Biol. and Health. New York: Springer

Vovk VG. 1993. A logic of probability, with application to the foundations of statistics (with discussion and a reply by the author). *J. Roy. Statist. Soc. Ser. B* 55:317–351

Wagenmakers EJ. 2007. A practical solution to the pervasive problems of p values. *Psychon. Bull. & Rev.* 14:779–804

Wakefield J. 2009. Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol.* 33:79–86

Wang X, George EI. 2007. Adaptive Bayesian criteria in variable selection for generalized linear models. *Stat. Sin.* 17:667–690

Wasserstein RL, Lazar NA. 2016. The ASA's statement on p-values: context, process, and purpose. *Am. Stat.* 70:129–133

Yuan Y, Johnson VE. 2008. Bayesian hypothesis tests using nonparametric statistics. *Stat. Sin.* 18:1185–1200

Zellner A. 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. PK Goel, A Zellner, vol. 6 of *Studies in Bayesian Econometrics and Statistics*, chap. 5. Amsterdam: North-Holland, 233–243

Appendix

A. SOME MATHEMATICAL RESULTS

A.1. The Folded Normal Distribution

A folded normal random variable $X \sim \text{FN}(\mu, \sigma^2)$ has density function

$$f(x) = \begin{cases} \frac{1}{\sigma}\left[\varphi\left(\frac{x-\mu}{\sigma}\right) + \varphi\left(\frac{x+\mu}{\sigma}\right)\right] & \text{if } x \geq 0 \\ 0 & \text{else.} \end{cases}$$

If X is normal, *i. e.* $X \sim \text{N}(\mu, \sigma^2)$, then $|X| \sim \text{FN}(\mu, \sigma^2)$.

A.2. Results for Sample-size Adjusted Bayes Factors

A.2.1. Convergence of the minimum Bayes factor based on the F-statistic to the "- e q log q" calibration. Here we show convergence of the minimum Bayes factor (24) based on the F-statistic to the "- e q log q" calibration (16) for $n = d + 3 \to \infty$. We first derive formula (25) for the minimum Bayes factor in the linear model if $n = d + 3$, so $f \sim F(d, 2)$ under H_0. In this case formula (24) simplifies (for $f > 1$) to

$$\text{minBF}(f) = \left(\frac{1 + 2/d}{f + 2/d}\right)^{(d+2)/2} f^{d/2} \tag{34}$$

and there is a closed-form expression for the F-statistic as a function of the P-value:

$$f = \frac{2}{d}\left[(1-p)^{-2/d} - 1\right]^{-1}. \tag{35}$$

Hence, $f > 1$ is equivalent to $p < 1-(1+2/d)^{-d/2}$ and that threshold converges from below to $1 - 1/e$ as $d \to \infty$. By plugging (35) into (34) and simplifying the expression, we find

$$\text{minBF}(f) = \underbrace{(d+2)\,\frac{1-(1-p)^{2/d}}{2}}_{\to -\log(1-p) \text{ for } d\to\infty} \underbrace{\left(\frac{d+2}{d}\right)^{d/2}}_{\to e \text{ for } d\to\infty} (1-p),$$

so we obtain

$$\lim_{d\to\infty} \text{minBF}(f) = -e\,(1-p)\,\log(1-p),$$

which is what we wanted to show.

A.2.2. Convergence of the Bayes factor based on the *F*-statistic to the Bayes factor based on the deviance. Here we show the convergence of the Bayes factor (23) based on the F-statistic to the Bayes factor (29) based on the deviance for $n \to \infty$.

Proof of claim (28). By assumption, the deviance z and $g > 0$ are fixed. First, we express the test-based Bayes factor $\text{BF}(f)$ (23) as a function of the deviance z instead of the F-statistic f by using relation (21) and the identity $R^2 = 1 - \exp(-z/n)$. This yields

$$\text{BF}(f) = (g+1)^{-(n-d-1)/2} \left[g \exp\left(-\frac{z}{n}\right) + 1\right]^{(n-1)/2}.$$

Rearranging the above formula gives

$$\lim_{n\to\infty} \text{BF}(f) = (g+1)^{d/2} \lim_{n\to\infty} \left[\frac{g \exp\left(-\frac{z}{n}\right) + 1}{g+1}\right]^{(n-1)/2}.$$

By using the series expansion of the exponential $\exp(-z/n)$ and the result $\lim_{n\to\infty}(1 + x/n)^{n-1} = \exp(x)$ for $x \in \mathbb{R}$, we then obtain the limit

$$\lim_{n\to\infty} \left[\frac{g \exp\left(-\frac{z}{n}\right) + 1}{g+1}\right]^{(n-1)/2} = \exp\left(-\frac{g}{g+1}\,\frac{z}{2}\right),$$

which completes the proof. □

A.3. Results for Large-sample Bayes Factors

A.3.1. Convergence of the minimum Bayes factor based on the deviance. Here we show convergence of the minimum Bayes factor (32) to the Edwards bound (12) for $d \to \infty$, adapting the proof from Held et al. (2015, Appendix B). The Edwards bound (12) is $\text{minBF}(t) = \exp(-t^2/2)$ with $t = t(p) = \Phi^{-1}(1-p)$ for any $p < 0.5$. For large d it is then sufficient to consider the case $z > d$, where the minimum Bayes factor (32) is

$$\text{minBF}(z) = \left(\frac{z}{d}\right)^{d/2} \exp\left(-\frac{z-d}{2}\right),$$

here z is the $(1-p)$-quantile of the chi-squared distribution with d degrees of freedom. We will show that for $d \to \infty$ and fixed P-value $p < 0.5$, the ratio $\text{minBF}(z)/\exp(-t^2/2)$ is 1.

First note that with $d \to \infty$, the standardized chi-squared distribution converges to a standard normal, so $(z-d)/\sqrt{2d} \overset{a}{\sim} \mathrm{N}(0,1)$ and hence $z \approx d + \sqrt{2d}\,t$. Plugging this into formula (32), we obtain

$$\begin{aligned}\frac{\text{minBF}(z)}{\exp(-t^2/2)} &\approx \left(\frac{d+\sqrt{2d}\,t}{d}\right)^{d/2} \exp\left(-\sqrt{\frac{d}{2}}\,t + t^2/2\right) \\ &= \exp\left[-at + a^2\log(1+t/a) + t^2/2\right]\end{aligned}$$

with $a = \sqrt{d/2}$. Now for large d the term t/a is small and, hence, we can apply a second-order Taylor expansion of $\log(1+x) \approx x - x^2/2$ around $x = 0$. This yields

$$\frac{\text{minBF}(z)}{\exp(-t^2/2)} \approx \exp\left[-at + a^2\left(\frac{t}{a} - \frac{t^2}{2a^2}\right) + \frac{t^2}{2}\right] = \exp(0) = 1,$$

which proves the statement.

A.3.2. Combining minimum Bayes factors. Here we prove the claim made in Section 4.3 that the product minimum Bayes factor is smaller than or equal to the combined minimum Bayes factor based on $z_+ = \sum_{i=1}^n z_i$ and establish when equality holds.

Note that the minimum Bayes factor (32) is obtained by minimizing the Bayes factor (29) with respect to g. We will thus start by considering the product and the combined Bayes factor based on (29). This product Bayes factor is

$$\begin{aligned}\prod_{i=1}^n \text{BF}(z_i) &= \prod_{i=1}^n \left[(g_i+1)^{d_i/2} \exp\left(-\frac{g_i}{g_i+1}\frac{z_i}{2}\right)\right] \\ &= \prod_{i=1}^n (g_i+1)^{d_i/2} \exp\left(-\sum_{i=1}^n \frac{g_i}{g_i+1}\frac{z_i}{2}\right) \qquad (36)\end{aligned}$$

and the combined Bayes factor based on z_+ with $d_+ = \sum_{i=1}^n d_i$ is

$$\text{BF}(z_+) = (g+1)^{d_+/2} \exp\left(-\frac{g}{g+1}\frac{z_+}{2}\right). \qquad (37)$$

For $g_i = g$ for all $i = 1, \ldots, n$, the product Bayes factor (36) is equal to the combined Bayes factor (37). To obtain the product minimum Bayes factor, each g_i in (36) is optimized separately to minimize the corresponding term for $i = 1, \ldots, n$. This leads to a minimum Bayes factor that does not exceed the combined minimum Bayes factor obtained by choosing g to minimize (36) under the restriction $g_i = g$ for all $i = 1, \ldots, n$. It follows that the product minimum Bayes factor cannot be larger than the combined minimum Bayes factor.

To see when equality holds, note that the estimates of g_i for the product minimum Bayes factor are $\hat{g}_i = \max\{z_i/d_i - 1, 0\}$ and the estimate of g for the combined minimum Bayes factor is $\hat{g} = \max\{z_+/d_+ - 1, 0\}$. So equality holds if all z_i (or equivalently all p_i) and all d_i are equal or if $z_i < d_i$ for all $i = 1, \ldots, n$.